Supplementary material for:

DNA Base Pair Resolution by Single Molecule Force Spectroscopy

By Bernie D. Sattin, Andrew E. Pelling, and M. Cynthia Goh

Data analysis:

(a) background

If there is a large number of independent variables, a normal distribution is expected, and previous analysis of force spectroscopy was thus performed by fitting the histogram of the data to a Gaussian, whose maximum was taken as the true value of the experimental data. Following Lee *et al.* (1), the sequence we used was expected to give rise to three distinct, but overlapping, Gaussian curves corresponding to 12, 16, and 20 bps (one for each of the major binding modes). While one can force the data to be fitted in this manner, examination of the histograms generated by our large data set (see Figure S1 for a representative set) made it clear that there were not three distinct peaks. Furthermore, the presence/absence of peaks as well as their location was highly variable *depending on the parameters used to create the histogram*. For instance, manipulating the start value or bin size can drastically affect the data. This can be demonstrated taking the following data set: 0.9, 1.0, 1.1, 1.4, 1.5, 1.6, 1.8, 1.9, 2.0. Visual inspection would appear to indicate 3 distinct groups, centered at 1.0, 1.5, and 1.9. But if one were to create histograms of bin size 0.1, and start counting at 0, one gets a single bin at all data points. Increasing bin size to 0.2, the bins collect differently. There is no objective criterion for choosing these parameters, and one cannot find a distribution that gives peaks at the appropriate locations and nowhere else. Thus, analysis by histograms and the subsequent fitting to a normal distribution, while qualitatively useful, actually introduces more uncertainty to the measurement. We decided to dispense with this method, and introduce a new approach – the use of cluster

analysis – to handle our data, which is qualitatively compatible with histogram results, but enabled a more objective assessment of the data.

In cluster analysis, data are grouped by similarity in proximity, as measured by Euclidean distance, into clusters of similar size. Thus, one forms highly homogeneous groups that highly heterogeneous when compared with each other. This similarity is represented as a dendritic tree, where the most dissimilar clusters are separated by the greatest distance. Additionally, the length of individual branches of the tree indicates the actual distance between two adjoining groups. We can create a dendrogram with the minimum cluster size being the Z resolution of our instrument (0.01 nm). The large size of our data set allows us to do this. The goodness of fit of a cluster dendrogram can also be ascertained. This cophenetic correlation function (ccf) allows one to determine at what level the tree becomes to dissimilar to be meaningful, with a value closer to unity indicating a better fit. The ccf can also be used to determine which clustering method gives the best dendrogram.

(b) application

Cluster analysis was performed on our force data. First, individual data sets were placed into order of increasing cantilever deflection during rupture (double dagger in Fig. 2) which, when multiplied by the spring constant, gives the force. Next, the data were clustered into the number of expected groups plus one. This extra group was added to allow for the large and multiple rupture events to be collected into the highest value cluster. The clustering was done by first measuring the Euclidean distance between pairs of objects in the data set, then arranging the distances in increasing order into a similarity matrix. This distance was computed four different ways: Euclidian, Mahalanobis, city block, and Minkowski. Next a hierarchical cluster tree was created from each of the four similarity matrices using five different measurements – shortest

distance, longest distance, average distance, centroid distance, and an incremental sum of squares. Finally, the set number of clusters from the hierarchical cluster tree was created. A ccf was used to determine the best fit dendrograms. In all cases, the Euclidian similarity matrix and average distance hierarchical tree produced the best fit for the data sets.

In order to reassure ourselves that the cluster analysis was predicting the correct values, we plotted the histograms beneath the clusters (Figure S1). As mentioned above, histograms containing multiple peaks are easily affected by changes in the grouping parameters. To address this issue, we placed histograms of increasing bin sizes on top of one another. This makes it easy to see the concentration of data in certain areas (an insight gained from a P.M. Williams paper (2)). It is clear from this picture that the end branches of the dendrograms correspond to points of high data density. The average value of the branch was then taken to represent the branch, and used in further data analysis. One may postulate that the analysis could be further validated by examining the relationship between the rupture force and the z-distance – i.e., the extension of the helix at pull off. Unfortunately, this analysis is not simple due to invalidated by the lack of knowledge of the precise geometry of the strands being separated. If there is a slight angle between the anchor of the ODN on the tip and the ODN on the surface, the rupture distance is decreased. This uncertainty is compounded by the fact that the precise length of the duplex DNA is unknown at the time of rupture (i.e. is it relaxed, or elongated, and by how much?), and that the duplexes of 16 and less base pairs can form in more than one configuration.

Supplemental References

1. Lee, G.U., Chrisey, L.A. and Colton, R.J. (1994) Direct Measurement of the Forces between Complementary Strands of DNA *Science*, **266,** 771-773.
2. Allen, S., Chen, X., Davies, J., Davies, M.C., Dawkes, A.C., Edwards, J.C., Roberts, C.J., Sefton, J., Tendler, S.J. and Williams, P.M. (1997) Detection of antigen-antibody binding events with the atomic force microscope *Biochemistry*, **36,** 7457-7463.

Figure caption

Figure S1. Thousands of force curves are collected, and their rupture forces are summarized as a histogram with values defined by cluster analysis. The histograms (shown at two different bin sizes) clearly have some clustering. The clusters are determined and assigned to increasingly larger numbers of bps. (A) TIP/MATCH interaction, representing 9 through 20 bps. (The 13^{th} cluster which contains rupture events from 5 to 15 nm is omitted in the interest of space.) (B) TIP/1MIS interaction, representing 9 through 19 bps. (The 12^{th} cluster which contains rupture events from 5 to 15 nm is omitted in the interest of space.). (C) TIP/2MIS interaction, representing 9 through 18 bps. (The 11^{th} cluster which contains rupture events from 5 to 15 nm is omitted in the interest of space.)

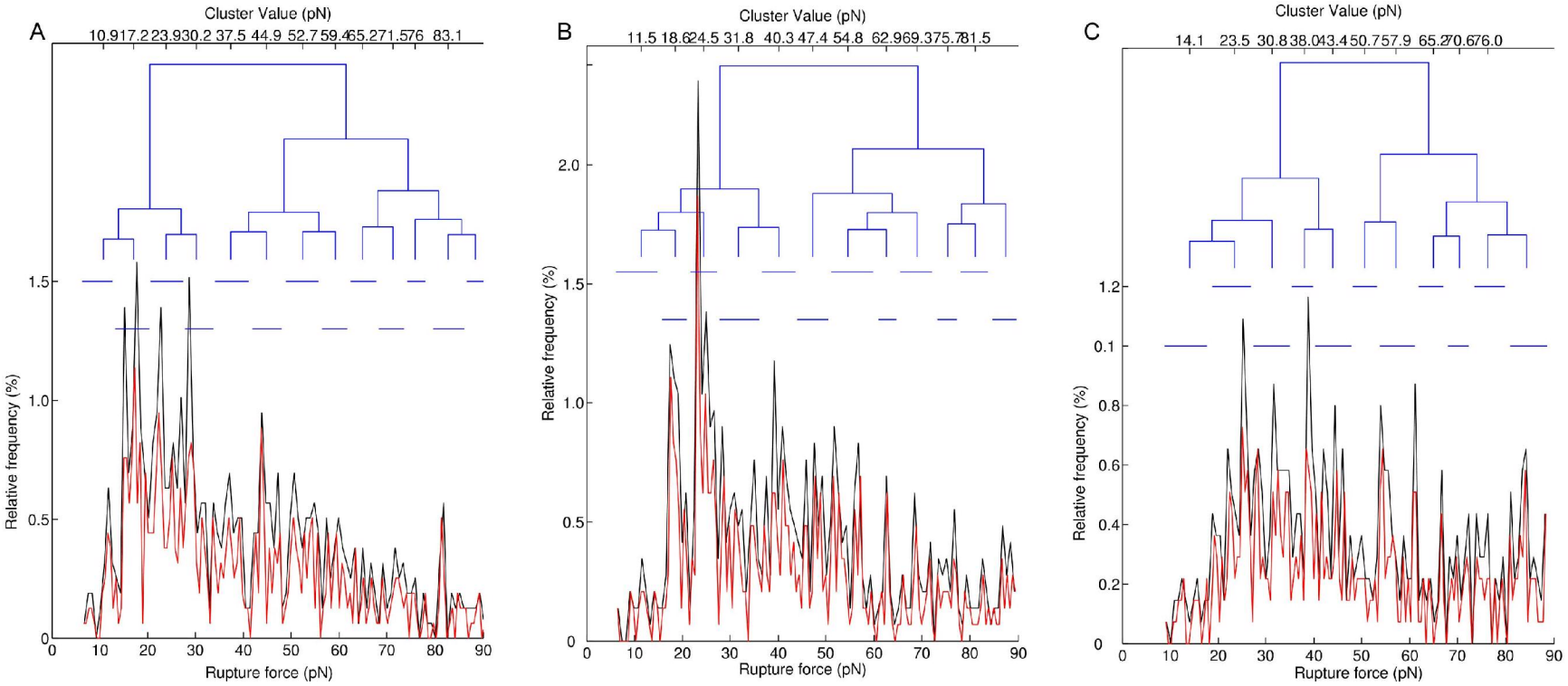

Figure S1. Thousands of force curves are collected, and their rupture forces are summarized as a histogram with values defined by cluster analysis. The histograms (shown at two different bin sizes) clearly have some clustering. The clusters are determined and assigned to increasingly larger numbers of bps. (A) TIP/MATCH interaction, representing 9 through 20 bps. (The 13th cluster which contains rupture events from 5 to 15 nm is omitted in the interest of space.) (B) TIP/1MIS interaction, representing 9 through 19 bps. (The 12th cluster which contains rupture events from 5 to 15 nm is omitted in the interest of space.). (C) TIP/2MIS interaction, representing 9 through 18 bps. (The 11th cluster which contains rupture events from 5 to 15 nm is omitted in the interest of space.)